PhD Research Proposal: Relating audio and 3D scenarios in audiovisual productions

David García Garzón

June 2, 2008

1 Introduction and motivation

Audiovisual productions, such as films or video games, combine two main elements: audio and video. In the last years, technologies for rendering video from synthetically designed 3D scenes have achieved a great degree of realism, both doing real-time and off-line rendering. Regarding audio, current productions have achieved a high degree of realism and are capable of making the audience feel totally immersed in the intended scenario. However, in contrast to video rendering, the audio production process is still quite a hand crafted job, lacking any automatism that uses such scene definition. Therefore, achieving a minimal level of realism presently demands a high amount of skills and effort to audio engineers.

To illustrate the lack of relation between audio and video in the course of a audiovisual production, let us consider the work-flow of a high-end film production. After the shooting stage, the audio engineer is given a set of audio tracks recorded on site, together with a low resolution version of the images shot. Even if the images have been synthesized with computer graphics technologies, all that the audio engineer receives is the rendered video, with no metadata whatsoever describing the geometry, nor the materials, nor the positions of the sound sources in the scene. Thus, one of the first tasks of the audio engineer is to figure out what sort of reverberation should the sounds have by staring at the video.

Later in the postproduction stage, the audio engineer needs to know in which sort of exhibition system(s) will the production be playbacked. In a typical Japanese production, such exhibition systems might easily be 5.1, 7.1 and even 22.2 [1] surround setups. This implies that a different editing and mixing must be done for all such options, obliging the audio engineer to carefully fine tune a large number of channels.

The described *state of affairs* certainly calls for research to provide solutions and automatization of the audiovisual work-flow. The general goal of the research project presented here is that of binding 3D and audio once and for all. That is, binding audio material of audiovisual productions and the scenario they are intended to be sounding in, so that most of the 3D audio rendering process can be done automatically and, when required, in real-time.

At a more detailed level, the research will focus on four aspects of this general goal:

- **Reverb Automatization, aka Audio Rendering:** capturing the acoustic properties of a 3D scenario

- **Physical coding:** coding the necessary magnitudes of the acoustic field that guarantees adaptability to any exhibition system.

- **3D decoding:** reproducing as faithfully as possible the acoustic field given any 2D or 3D exhibition system

- **Environment acoustic inference:** inferring a plausible scenario given an audio recording

2 State of the art

This section describes the state of the art of the several fields related to this research. Such state of the are will be explained fitting the four different aspects we are addressing in this research as enumerated in the previous section.

2.1 Audio rendering: Capturing 3D scene acoustics

In order to capture how the scenario modifies the perception of the sound, the sound wave must be simulated considering its interactions with the scenario's geometry and materials. From such a simulation one should expect obtaining the same measures that would be obtained by a microphone placed at the listener position within the virtual scenario.

Under normal levels of sound pressure, all transformations over the wave can be considered linear. That means that the way an scenario modifies a wave from a source point to a target point can be fully characterized by an impulse response. An impulse response is a signal that has a flat frequency distribution, so recording or simulating how it propagates from the source point to the listener point contains information on how each frequency is modified. By having just the impulse response, we can obtain the response of any other sound by using the impulse response as a filter, that is convolving the emitted signal with the impulse response. Note that an impulse response is tied to a given source and listener positions.

In summary, the first problem to solve can be formalized as follows: Finding the impulse response that characterizes how the signal is transformed from a source position to a listener position while traversing the scenario. The literature explores two main families of algorithms that pursue such a goal:

- Wave-based methods
- Geometric methods

Wave-based methods compute a discrete approximation of the solution to the wave equation, which is the equation that describes the sound propagation through an environment [2]. Several wave-based methods exists such as Finite Elements Method (FEM), Boundary Elements Method (BEM) [3], and Finite-Difference Time Domain (FDTD) [4]. Although all of them are computationally expensive, the growing computing capabilities of standard computers have made them an interesting path to take.

Geometric methods more widely used because they are cheaper computationally. They exploit the fact that light and sound propagation share a number of the similarities in order to use a number of techniques that have been using in computer graphics to visually render 3D scenes. Of course, such techniques as is do not take into account some phenomena that can be neglected with light such as diffraction and interference, so computer graphics methods must be modified to consider them. Commonly used techniques in this family are Image Source [5], Ray Tracing[6] and Beam Tracing [7].

2.2 Physical encoding

In order to reproduce the acoustic field as closely as possible to a real situation, it is not enough to encode only the pressure signal at one point. The reason is that the pressure field is omnidirectional, it does not contain any information about the direction of the incoming waves.

Ambisonics [8] was the first technology to recognize this fact, and to propose the recording of other magnitudes beyond the pressure signal. The principle behind this technology is to decompose the pressure of the acoustic field at one point into spherical harmonics, and to record the time varying coefficients of such harmonics. In particular, the coefficients corresponding to the first spherical harmonic coincides with the pressure signal at that point. Similarly, the coefficients corresponding to the next three spherical harmonics coincide with the three components of the velocity vector of the air fluid at that point.

Ambisonics systems based on first four spherical harmonics (pressure and velocity) are called First Order Ambisonics. Those using harmonics beyond these are named Higher Order Ambisonics (HOA).

The extra information contained in harmonics beyond the pressure signal can be used in the reproduction stage to create a more faithful acoustic field capable of localizing sounds (perceiving them from the right direction).

As it happens with any expansions of a function in a given basis, the more coefficients are encoded, the less error is made in extrapolating the field beyond the recording point [9]. This implies that HOA typically exhibit a larger sweet spot (region where the sound field is reproduced correctly) than First Order Ambisonics.

FAO has been implemented successfully in a professional microphone (Soundfile [10]), widely used nowadays in broadcasting and music recordings. However HOA still lacks of the corresponding microphone due to the larger number of channels that need to be recorded.

2.3 Decoding to exhibitions systems

Exhibition systems for localized audio can be classified into *binaural technologies* and *loudspeaker arrays technologies*. Binaural technologies are the ones that reproduce the sound as it were sounding at each ear. They use earphones to reproduce the sound so two channels are decoded. Loudspeaker arrays are systems that use sets of many loudspeakers placed arround the listener to reproduce the desired wave field. They normally decode as many channels as loudspeakers.

2.3.1 Binaural technology

Binaural technologies encode into two channels several of the cues that the brain actually uses to locate sound sources using the sound perceived by each ear. Such cues are then fed directly to the earphones as the sound as they were arriving from an external wavefield.

The brain uses several cues to determine which is the source event localization [11]. The most important ones are Interaural Time Differences (ITD) for the low frequencies and Interaural Level Differences (ILD) for the high frequencies [12]. Low frequencies can diffract around the head and reach both ears whichever the direction they come from, but they reach each ear with a different delay. The brain is able to detect such delay and triangulate the position. Such triangulation is harder for higher frequencies as the time delay goes beyond the wave period. But high frequencies are not able to diffract around the head and, thus, the head filter out such frequencies and different sound levels reach each ear. A simple decoding technique for binaural synthesis consists on appling ILD and ITD to the incoming audio.

Interaural differences are important cues but not the only ones. Sound events located within the axial plain, that is equidistant to both ears, can still be successfully located by our brain. One of the most important cues our brain uses in that case is the different filtering that the pinnae, the external ear, does to the sound depending on the incoming direction.

Most cues, such as ITD, ILD, pinna filtering, torso reflections... can be captured in a Head Related Impulse Response (HRIR), that is the response to an impulse signal coming from a given direction referred to the head. They are often named by their frequency domain equivalent: the Head Related Transfer Response (HRTF). HRTF's databases are obtained by sampling HRTF functions at different directions. So a more sophisticated method to decode binaural channels consists in convolving the incoming audio with the HRTF functions which is closer to the direction the sound source is relative to the listener.

There several methods to obtain HRTF databases. Some databases, such as the one made available by the MIT [13], are measured using a manikin. Some others, such as the ones offered by the IRCAM [14], have been recorded by inserting microphones on the ears of real subjects. Due to the costs of doing the meassures, lately some efforts are driven towards the simulation of HRTF functions using a geometrical 3D model of the subject [15]. In any case, HRTF's are very dependant on the subject, so that an HRTF database measured for a given subject, can lead to bad localization results used with a different subject. Some studies have been done on relating HRTF's to subject's anthropometric measures, but they are very limited both in the extension of the analysis and the applicability of the results [16].

The brain uses more cues than enforce the main ones. One of them, could be called audio parallax in a analogy to visual parallax. Visual parallax is a sensation of depth we get when objects in the scene change their projection differently on the cornea due to a change of position of the object or the viewer. By analogy, head movements or sound source movements create differences on localization cues that can help to better detect them. Some experiments [17] confirm that adding head tracking to modify the HRTF accordantly enhances the localization of sources.

Also visual cues are important to enforce sound event localization cues, so that they can void any acoustical cue. This is why listening experiments should avoid providing any visual cue that can distort the results [11].

Direct convolution of the sound with the HRTF just works when coding direct sound. When having an Ambisonics representation of the signal, such the one we are obtaining from the audio rendering, other methods should apply. Higgins proposes an computational optimal method [18] which converts the database into a set of equivalent HRTF functions to be convoluted with each Ambisonics component.

2.3.2 Loudspeaker arrays (surround systems)

Until 1931, all loudspeakers reproduction systems were essentially mono. Even if more than one loudspeaker was used, they share the same single audio channel. Blumlein invented and patented stereo technology [19], which is capable of locating sound sources within a 60 degrees range in front of the listener. It was not until the 90s that systems with a higher number of channels were standardised, mainly thanks to the incorporation of 5.1 systems in cinema theaters. These provided a reasonably good localization within the whole horizontal plane of the listener, despite some deficiencies in the back, mostly due to the lack of speakers in that range.

The present situation is that new exhibition systems with more loudspeakers are competing in the market. Some of them include the ability to place sound source out of the horizontal plane by means of placing loudspeakers at different height levels. The latest one, at the moment, is the 22.2 system in standardisation process by the NHK in Japan [1], [20], [21].

The state of the art regarding production of audio material for such systems is twofold. On the one hand, most such productions have inherited the techniques from stereo and 5.1: essentially simple amplitude panning. Indeed, hardly any production exploits the full possibilities of surround sound, as most of the audio is playbacked through the frontal loudspeakers due to constraints such as images being presented always in front of the audience.

On the other hand, recent efforts have focused in exploiting the Ambisonics technique described above. At the decoding stage, the problem is to find the signals to be fed to the loudspeakers of a given reproduction system, such that the spherical harmonics of the pressure field at the listener position is as close as possible to those recorded or computed at the encoding stage. This is a problem that requires both physical and psycho-acoustical considerations. The goal is typically to define a suitable cost function the minimization of which leads to a compromise among all such consideration [9], [22]. Unfortunately the cost functions are typically highly non-linear, leading to a search space full of local minima.

Before concluding this section, it is worth pointing out that the last decade has witnessed the irruption of an alternative exhibition system named Wave Field Synthesis (WFS). Based on the exploitation of the Huygens principle, it theoretically allows the reproduction of the exact acoustic field within the complete listening space. However we will not research into this topic due to the fact that WFS requires such a large number of loudspeakers (typically above 100) that it constraints its application to very special high-end events. For more information see [23].

2.4 3D Acoustics Inference

The problem of 3D Acoustics Inference (3D-AI) consist in finding a 3D environment which is able to cast the same acoustic properties that a given recorded audio. That is still a new field of research. The closer research that has been undertaken is possibly the topic of inferring source localization given some kind of recording. Two approaches have led to rather successful results:

- emulating human localization mechanisms on binaural recordings [24],
- recording the acoustic field with multiple microphones to deduce the source location from the different delays on the arrival time [25].

In any case, to the best of our knowledge, there is no literature on the more complex problem of 3D-AI.

3 Proposal

The research will focus on the four aspects of the audio workflow mentioned in section 1.

3.1 Audio Rendering

Concerning audio rendering, the main goal of this research is enhancing the physical accuracy and the speed of existing methods. At this point, several opportunities can be foreseen. One could be enhancing the physical accuracy of geometric models by considering effects such as diffraction and interference, typically out of reach of the geometrical approximation. Another possibility would be building hybrid algorithms by combining geometrical methods and wave-based ones such as FDTD to enhance the accuracy on lower frequencies of the former. The problems to solve in hybrid models is to find out the band in which each simulation should be valid, how to mix both results and how to speed up the wave-based methods using that band limitation.

3.2 Physical encoding

As mentioned in the state of the art, Ambisonics technology offers a novel and powerful way to capture acoustic fields. While it has been used extensively used in microphone technology for recording real events, it has hardly penetrated into the field of acoustic simulations for audiovisual productions. First Order Ambisonics (FOA) is currently being implemented into simulations by the Audio Group at Barcelona Media, producing rather successful initial results.

However, it is conceivable that the major advantage of incorporating Ambisonics to simulations will come from Higher Order Ambisonics (HOA). The benefit of using HOA components is the widening of the sweet spot and the increase of localization accuracy. Note that virtually all problems found in incorporating HOA to microphones are not present in computer simulations. Essentially, when placing a large number of microphones almost coincidently, they shadow each other. Another problem is that microphones have directivity patterns that vary with frequency, spoiling the principles of Ambisonics beyond 8-9KHz. Of course those mechanical problems will be absent in our simulations.

We plan to apply Ambisonics encoding to the reverberation part of the audio but not necessarily to the direct sound. The reason is that in many applications, localization is of crucial relevance (e.g. video games), and Ambisonics, until certain high order, spread the localization of the sources. Direct sound and early reflections can be processed directly using HRTF's or a simple panning algorithm.

Direct sound and the early reflections can be computed faster than the reverberation cue that needs more rebounds in a geometric simulation. On the other hand, early rebounds and direct sound provide the information to localize the sound so they need frequent updates. The reverberant cue is more stochastic and it provides more information about the scenario itself than location. It does not change that much with movements of the listener or the source unless the enclosing room changes. So a possible optimization of the process could be separating the codification into the reverberant cue component and the direct sound plus early reflections at different update rates.

3.3 Exhibition system decoding

Regarding the decoding, we will proceed in three fronts:

Algorithm design. We will study and design signal processing algorithms to decode the three-dimensional surround recordings into sets of signals to feed the loudspeaker systems. These algorithms should be flexible enough to provide optimal decoding of the surround signal sets to any desired 3D exhibition system and they should be downwards compatible with existing two-dimensional setups. Of course, this research must be based on physical principles and the algorithms have to be validated by means of psychoacoustic tests. It is important for these algorithms to provide virtually any number of output channels. The idea is that once the Ambisonics components are correctly encoded with a few channels, these are stored and delivered to the end user, whose equipment will perform the optimal decoding to match the listening setup. This approach will overcome the problems related to the preparation, storage and distribution of many different formats, and it will surmount the limitations of media and bandwidth, allowing the whole information to be easily stored on optical discs or streamed over the web and subsequently decoded.

Binaural decodification Still a lot of enhancements can be done on binaural decoding. While the algorithm to compute the Ambisonics equivalent HRTFs supposes an infinitesimal and uniformly distributed set of HRTFs, most available HRTF database are not complete enough or not dense enough for the variation to be neglected. A typical example is that most measured databases are missing the lower elevations. Also, using ambisonics HRTF equivalents enables having a finer resolution than the one of the HRTF's, to do head tracking of smaller head movements.

Speaker layout design. Here we plan to answer the question of what is the ideal optimal 3D reproduction system. Of course the answer may not be unique and perhaps it will depend on the purposes of the system. Basing on physical and psychoacoustic principles, the goal is to design a loudspeaker configuration to offer an accurate reconstruction of the three-dimensional soundfield for improved realism. At the same time we will aim at extending the dimension of the sweet spot, to allow a larger number of listeners to correctly perceive the sound.

3.4 3D Acoustics Inference

The last aspect of the research is 3D Acoustics Inference. As stated above this is a new research line in the literature. We see a clear plan of research to be undertaken and exploited and several nice applications that such technology will provide.

We will use our room acoustics simulation capabilities to define a parameter space (for example, room dimensions, absorption coefficients...) that determines the acoustics of the scene. Given a recorded audio track, we will define a cost function that defines a metric in this parameter space, such that the distance is zero when the simulated acoustics match those of the recording. By minimizing such a function we will determine the scene that better suits that audio.

Research will be undertaken to investigate the properties of this procedure. For example: are there many local minima? what do they physically correspond to? are shoe-box-shaped geometries generic enough to fit most of the acoustics? if not, how much should we complicate geometries? in other words, how large should the parameter space be?

We shall also note that the goal is not necessarily that of having an identical scenario to the one present in the recording. For most applications inferring one that sounds indistinguishably close to the real one will be more than enough.

We envisage a number of interesting applications:

- An audio engineer wants to introduce dubbed dialogs in an recorded scene that match the acoustics properties of those recorded during shooting.

- An audio engineer wants to modify an existing reverberation preset using high level room parameters (room size, wall materials) instead of tweaking low level impulse response parameters.

- An audio engineer has a found a reverb that he wants to apply to dubbed dialogs and effects, that are supposed to take place in a reception at the US embassy. He uses this technology to obtain a plausible room, locates the sound sources within it and navigates freely through it hearing the change of the localization of the sources accordingly.

- An audio engineer wants to change the localization of some sound event in a production but the scenario is not available.

- An audio engineer wants to change the listener position from the one the microphone was to actual point of view of the final production.

References

[1] K. Hamasaki, K. Hiyama, and R. Okumura, "The 22.2 Multichannel Sound System and Its Application," in *AES 118th Convention*, 2005.

[2] T. J. Chung, *Computational Fluid Dynamics.* Cambridge University Press, 2002.

[3] A. F. Seybert, C. Y. R. Cheng, and T. W. Wu, "The solution of coupled interior/exterior acoustic problems using the boundary element method," *Journal of the Acoustic Society of America*, vol. 88, no. 3, pp. 1612–1618, September 1990.

[4] D. Botteldoren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *Acoustical Society of America*, vol. 98(6), 1995.

[5] H. Nironen, "Diffuse reflections in room acoustics modelling," Master's thesis, Helsinki University of Technology, 2004.

[6] A. Farina, "Pyramid tracing vs. ray tracing for the simulation of sound propagation in large rooms," *Proceedings of International Conference on Computer Acoustics and its Environmental Applications*, 1995.

[7] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, and J. West, "Modeling sound reflection and diffraction in architectural environments with beam tracing," *(invited paper) Forum Acusticum*, September 2002.

[8] "Ambisonics.net." [Online]. Available: http://www.ambisonics.net

[9] M.A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio Eng. Soc.*, vol. 33(11), pp. 859–871, Nov. 1985.

[10] "Soundfield.com." [Online]. Available: http://www.soundfield.com

[11] J. Blauert, *Spatial hearing, the psychophysics of human sound localization.* MIT Press, 1997.

[12] Lord Rayleigh, "On our perception of sound direction," *Phil. Mag.*, vol. 13, pp. 214–232, 1907.

[13] B. Gardner and K. Martin, "HRTF Meassurements of a KEMAR Dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep. 280, 1994.

[14] "Listen HRTF Database." [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[15] W. Kreuzer and Z. Chen, "A fast multipole boundary element method for calculating hrtfs," in *AES Convention*, 2007.

[16] Patrick Satarzadeh1 and V. Ralph Algazi1 and Richard O. Duda1, "Physical and Filter Pinna Models Based on Anthropometry," in *AES Convention*, 2007.

[17] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "The importance of head movements for binaural room synthesis," in *Proceedings of the 2001 International Conference on Auditory Display*, July 2001.

[18] B. Wiggins, "An investication into the real-time manipulation and control of three-dimmensional sound fields," Ph.D. dissertation, Univertity of Derby, 2004.

[19] Alexander, Robert Charles, *The Inventor of Stereo: The Life and Works of Alan Dower Blumlein.* Focal Press, 1999.

[20] K. Hamasaki, S. Komiyama, K. Hiyama, and H. Okubo, "5.1 and 22.2 Multichannel Sound Productions Using an Integrated Surround Sound Panning System," in *Proceedings NAB BEC*, 2005.

[21] K Hamasaki and K Hiyama and T Nshiguchi and K Ono, "Advanced multichannel audio systems with superior impression of presence and reality," in *AES 116th Convention*, 2004, convention paper.

[22] David Moore and Jonathan Wakefield, "The Design of Ambisonic Decoders for the ITU 5.1 Layout with Even Performance Characteristics," in *124th AES convention*, no. 7473, 2008.

[23] Jérôme Daniel and Rozenn Nicol and Sébastien Moreau, "Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging," in *114th AES convention*, 2003.

[24] F. Keyrouz and K. Diepold, "A novel biologically inspired neural network solution for robotic 3d sound source sensing," *Soft Comput*, vol. 12, p. 721–729, 2008.

[25] Trinnov Audio, "5.0 Sound recording in High Spatial Resolution." [Online]. Available: http://www.trinnov.com/download_file.php?file=Trinnov_SRP_GB.pdf